*Economics and Philosophy*, **27** (2011) 273–295 Copyright © Cambridge University Press
doi:10.1017/S0266267111000216

# TRANSCENDENTAL ARGUMENTS AND INTERPERSONAL UTILITY COMPARISONS

**MAURO ROSSI**
*Université du Québec à Montréal, Canada*
*redsmau@gmail.com*

According to the orthodox view, it is impossible to know how different people's preferences compare in terms of strength and whether they are interpersonally comparable at all. Against the orthodox view, Donald Davidson (1986, 2004) argues that the interpersonal comparability of preferences is a necessary condition for the correct interpretation of other people's behaviour. In this paper I claim that, as originally stated, Davidson's argument does not succeed because it is vulnerable to several objections, including Barry Stroud's (1968) objection against all transcendental arguments of a 'strong' kind. However, I argue that Davidson's strategy can still achieve results of anti-sceptical significance. If we reformulate Davidson's argument as a 'modest' transcendental argument and if we embrace an 'internal' account of epistemic justification, it is in fact possible to have at least justified beliefs about how different people's preferences compare in terms of strength and about their interpersonal comparability.

## 1. INTRODUCTION

There are few solutions to the problem of interpersonal comparisons (ICs henceforth) of utility that have received less attention in the relevant literature than Davidson's proposal (Davidson 1986, 2004). This has certainly to do with the unfamiliar character of Davidson's philosophy for the typical audience concerned with the problem of ICs, which, amongst

I would like to thank Richard Bradley, François Maniquet, Dominic Martin, Christine Tappolet and two anonymous referees for their comments on previous drafts of the article.



https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

others, includes economists and social scientists. This lack of attention is however unfortunate. While, as I shall argue below, Davidson's argument does not succeed as originally stated, it can achieve some interesting results of anti-sceptical significance when suitably modified. The goal of my paper is to explain how and why. The structure will be the following. In section 2, I shall illustrate in more detail what the problem of ICs of utility is. In section 3, I shall present Davidson's solution. In section 4, I shall reconstruct the general form of Davidson's argument and discuss some objections against it. In section 5, I shall formulate and defend a modified version of his argument. In section 6, I shall examine the implications that this proposal has for the problem of ICs of utility.

## 2. THE PROBLEM

Following contemporary economic theory, I take utility to be a numerical representation of preferences. In this sense, the problem of ICs of utility is the problem of ICs of preferences. More specifically, I shall focus on the problem of comparing the *intensity* of different individuals' preferences. By 'intensity' I do not mean a phenomenological property that preferences might have, i.e. something that the individual having preferences subjectively 'feels' or 'experiences'. Rather, by 'intensity' I refer to a causal property of preferences, i.e. a property that leads the individual having preferences to act in specific ways in certain circumstances. ICs of preference strength so defined are often treated with scepticism in economics and other social sciences (cf. Robbins 1932; Arrow 1963; Hausman 1995). Indeed, the sceptic typically raises two sorts of doubts. (1) It is impossible to know how different people's preferences compare in terms of strength. (2) It is impossible to know whether or not preferences are interpersonally comparable in terms of strength. Let me explain.

If an individual's preferences satisfy certain technical axioms, they can be represented by an interval utility function, unique up to a positive affine transformation (von Neumann and Morgenstern 1944).[1] In order to fix an interval scale of measurement, it is necessary and sufficient to fix two points, namely, the zero and the unit. In the case of preferences, the unit and the zero are typically fixed in correspondence with the options that the individual prefers, respectively, the most and the least, in a given preference domain. Fixing these points is what makes an individual's preferences *measurable* on an interval scale. In order for different individuals' preferences to be *comparable*, the scale

[1] I assume the framework of vNM utility functions for illustrative purposes only. It is worth keeping in mind that Davidson's proposal (and its revised version discussed in this paper) applies more generally to the representation of preference strengths through other utility functions.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

of measurement on which their preferences are represented must be a common one. This requires that the unit and the zero be fixed in correspondence with two options with respect to which the individuals' preferences have the same intensity. The problem is that there is no point where different people's preferences can be said to possess the same intensity with certainty.[2] Hence, there is no way of telling with certainty whether or not different individuals' preferences are effectively measured on a common interval scale. This leads the sceptic to her two claims. First, it is impossible to know *which* positive comparative relation holds between different individuals' preferences when they are compared in terms of strength. That is, it is impossible to know whether one individual's preferences are either stronger than, or as strong as, or less strong than another individual's preferences. Second, it is impossible to know *whether* any such positive comparative relation holds between different individuals' preferences. If one individual's preferences are neither stronger than, nor as strong as, nor weaker than another individual's preferences, then preferences are interpersonally incomparable in terms of strength. Thus, according to the sceptic, not only is it impossible to know how different people's preferences compare in terms of strength; it is also impossible to know whether they are interpersonally comparable in terms of strength in this sense.

One might suggest that the individuals' most preferred and least preferred options provide the common points that we need in order to represent different people's preferences on a common interval scale. The idea is quite natural: if the best and worst options provide the unit and the zero for representing each individual's preferences, why should we not suppose that they provide also a *common* unit and zero for representing different individuals' preferences? These options seem to play the same causal role in these individuals' minds. Thus, if intensity is a causal property of preferences, it seems plausible to assume that different individuals' preferences have the same intensity in correspondence with these options. This proposal is typically known as the 'zero-one rule' for making ICs, because it assigns the utility value 0 to the option at the bottom of each individual's preference ranking and the utility value 1 to the option at the top of each individual's ranking (Isbell 1959).

Most people reject the 'zero-one rule' either by claiming that it is a normatively inadequate method for making ICs (e.g. Sen 1970: 98; Rawls 1971: 323; Hammond 1991: 216) or by arguing that such a rule is unable

[2] This contrasts with the case of beliefs, which are both measurable and interpersonally comparable. Interpersonal comparability is guaranteed by the fact that the degrees of belief of different rational individuals coincide in correspondence with the necessary proposition, which fixes the unit of the belief scale, and the impossible proposition, which fixes the zero of the belief scale (see Bradley 2007).

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

to capture hidden differences in sensitivity between different individuals' preferences (e.g. Griffin 1986: 120; Binmore 2009: 550). However, these are the *wrong* reasons for rejecting the 'zero-one rule'. First, the problem of the normative role of ICs of preferences is entirely separate from the problem of whether or not ICs of preferences are epistemically possible. So, an unsatisfactory solution to the former problem can still be a satisfactory solution to the latter problem. Second, appealing to hidden interpersonal phenomenological differences effectively means rejecting the functionalist characterization of preferences that has been dominating decision theory since its foundations (Ramsey 1926). This requires an explanation, in the absence of which the objection is irrelevant.

The 'zero-one rule' for making ICs of preference strength should still be rejected, but for a different reason. Indeed, there seems to be no way to know whether the best and worst options in different individuals' preference rankings play identical causal roles. These options are the 'best' and 'worst' only relative to a given preference domain. But this preference domain is to some extent arbitrary. We ascribe preference strengths to explain and predict people's behaviour in various counterfactual situations. In order to do this, it is necessary to consider various potential options, in addition to those included in the initial preference domain. Once this is recognized, it becomes clear that the 'best' and 'worst' options in the initial preference domain may share no interpersonally common causal properties. If there were an 'absolute' maximum, i.e. an option that is preferred to any other options, and an 'absolute' minimum, i.e. an option that is dis-preferred to all other options, then preferences over these options would play the same causal role in different individuals' minds and could be taken to have the same intensity with certainty. In other words, if preferences were bounded, then they would also be interpersonally comparable (cf. Hausman 1995; Weintraub 1995). However, the problem is precisely that preferences seem to be *unbounded*. It is in fact always possible to imagine options that are, respectively, a little bit better and a little bit worse than the 'best' and 'worst' options in a given preference domain (cf. Bradley 2007: 97–98). If this is true, the latter options cannot be taken as common units and zeros for representing preferences on an interpersonally common interval scale. While they provide the points required for the measurability of each individual's preferences, they do not provide the points required for their interpersonal comparability.[3]

[3] An alternative way of reaching the same conclusion consists in assuming, unlike the previous discussion, that the initial preference domain is infinite or, more exactly, that the preference relation on the set of alternatives does not admit of either an upper or a lower limit. If this is the case, it follows once again that the options that are used to fix the unit

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

Before proceeding, let me add a word of clarification about the relationship between the sceptic's two claims. Although these are often treated together in the literature, it is important to recognize their distinct character. It should be clear that the sceptic's second challenge is logically prior. If it is impossible to know *whether* preferences are interpersonally comparable in terms of strength, it is a fortiori impossible to know *how* different individuals' preferences compare in terms of strength. On the other hand, the first challenge enjoys a relative autonomy. Even if we know that preferences are interpersonally comparable in terms of strength, it does not follow that the sceptic's first challenge is thereby automatically met. It might well be the case that, while we have some conclusive reason to think that there do exist common units and zeros, which make different individuals' preferences comparable in terms of strength, it is impossible to know in correspondence with exactly which options these common points are situated. Thus, it may still be impossible to have knowledge of specific ICs of preference strength.[4]

## 3. DAVIDSON'S PROPOSAL

Davidson's proposal (1986, 2004) targets primarily the second issue about ICs of preference strength, but, as we shall see below, it has obvious

and the zero in each individual's utility scale are not 'absolute' maxima or minima. Thus, they cannot be taken as interpersonally common units and zeros.

[4] Let me illustrate the relevance of this point in relation to the previous discussion. Consider the following proposal to save the 'zero-one' rule. One may claim that it is possible to fix a common unit and a common zero if we consider the set of all *conceivable* options. The idea is that different individuals possess identical preference strengths in correspondence with the options that they prefer, respectively, the most and the least within the maximal preference domain, i.e. the preference domain that includes all the options that the individuals can conceivably face. The options at the top and the bottom of the individuals' rankings are good candidates for being common units and zeros, respectively, because they play the same causal role in different individuals' minds. If this solution is tenable, then it is indeed possible to show that preferences are interpersonally comparable in terms of strength. Clearly, however, this solution does not guarantee that the absolute best and absolute worst options are easily identifiable, from an epistemic point of view. So, even if we grant the existence of such points, it may still be impossible to know how different people's preferences compare in terms of strength. This proposal provides at most a solution to the second challenge, but not to the first. On the positive side, *if* this solution is correct, *then* at least one of the sceptic's challenges is met. The problem is that, as some authors have argued (see, for instance, Bradley 2007, who also proposes an interesting, alternative solution to the problem of ICs), the idea of a maximal preference domain of the sort just illustrated seems to be inconsistent with the unbounded character of preferences. In fact, if we accept the idea that preferences are unbounded, then it is a conceptual truth that preferences do not admit of either an upper or a lower bound. If this is true, the previous solution does not work for the sceptic's second challenge either. For reasons of space, I shall ignore this debate here. In what follows, I shall attempt instead to offer a solution both to the first and to the second challenges that does not presuppose the existence of an absolute maximum and an absolute minimum.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

implications for the first issue as well. As a starting point of his analysis, Davidson characterizes the framework in which the problem of ICs arises as a sequence of three distinct steps, involving:

1. the determination of each individual's preferences and their representation through a (family of) utility function(s);
2. the interpersonal comparison of the individuals' utilities;
3. the formulation of a normative judgement (e.g. a distributive decision) based on such comparisons.

In this framework, the problem of ICs arises in the second stage. One needs to find a 'basis' for *comparing* preferences in addition to the 'basis' used for *determining* those preferences in the first stage. The problem is that there seems to be no such interpersonal basis. Davidson claims that this characterization is the very source of the problem and should be abandoned. What are the alternatives? One possibility consists in denying that ICs are independent factual statements and maintaining instead that they are either normative or evaluative judgements. This position suggests that the second and the third steps are much more interdependent and difficult to distinguish than in the original framework, up to the point where they mesh together. If this is the case, the basis for ICs of preferences is normative (cf. Robbins 1932; Jeffrey 1971; Schick 1971; Hammond 1991). Another possibility consists in emphasizing the interdependence existing between the first and the second steps. Davidson pursues this line. More specifically, Davidson claims that the very attribution of preferences to another individual involves a sort of interpersonal comparison between the interpreter's and the agent's preferences. One does not have to find a basis for comparing preferences interpersonally in the second stage, *after* the attribution of preferences. The basis that assures the interpersonal comparability of preferences is given by the principle that makes preference ascription possible in the first stage, namely, the Principle of Charity (PoC henceforth).

Let us summarize Davidson's rather complex argument. The attribution of mental states to another agent is part of a process of interpretation of that agent's behaviour. Interpretation is a rational explanation of the agent's behaviour in terms of that agent's propositional attitudes (Davidson 1984). The goal of the interpreter is thus to explain the agent's behaviour in terms of mental states that have a specific function, content and strength. According to Davidson, correct interpretation is necessarily based on the PoC. This means that it is not possible to correctly interpret another agent's behaviour without employing the PoC. Roughly speaking, the PoC favours the interpretation that makes the interpreted agent's beliefs as true as possible, by the interpreter's own light. Crucially, the PoC urges the interpreter to project herself onto the interpreted agent's

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

mind. The interpreter projects her own mental concepts to characterize the functional properties of the other agent's mental states. She projects her own standards of truth to individuate the content of the agent's doxastic attitudes and her own standards of value to individuate the content of the agent's evaluative attitudes. Finally, the interpreter projects her own standards of rationality to identify the relationships between the agent's mental states.

According to Davidson, the interpreter's projection establishes a sort of *comparison* between her mental states and the interpreted agent's ones. As Davidson makes clear, this does not mean that 'the attributer consciously or unconsciously makes a comparison', but only that 'in the process of attribution the attributer necessarily uses his own values in a way that provides a basis for comparison' (Davidson 2004: 67). In other words, the interpreter's projection does not involve making any explicit IC *judgements*, but provides the *basis* for making such judgements at a later stage. Indeed, the interpreter's projection grounds the ascription of similarities and differences between the interpreter's and the agent's mental states across all the dimensions required for the interpretation of the agent's behaviour. The projection endows the agent with preferences that have the same general characteristics as the interpreter's own preferences. Thus, preferences play the same causal role in both individuals' minds. They possess a dimension of strength that has the same general features. On the other hand, similarities at the general level remain compatible with differences at the particular level. This means that the interpreter can attribute irrational preferences to the agent, when they do not satisfy her standards of rationality. Moreover, she can assign a propositional content to the agent's preferences that may differ from the content of her own preferences. Importantly, the interpreter can ascribe preferences to the agent that have a different intensity from her own preferences. According to Davidson, differences between interpreter and interpreted agent can be tolerated up to the point where the agent's behaviour remains *intelligible*, that is, up to the point where they serve the purpose of providing an intelligible explanation of the agent's behaviour. Yet, if the correct interpretation of the agent's behaviour requires establishing a comparison between different individuals' mental states, then preferences appear to be interpersonally comparable from the start. This is indeed Davidson's insight: correct interpretability *implies* interpersonal comparability. We do not have to ask ourselves whether preferences are interpersonally comparable in terms of strength in the second stage of the original framework. The very possibility of correctly attributing preference strengths to other individuals in the first stage implies that they are already interpersonally comparable.

Let us make Davidson's conclusion even more explicit by reconsidering the sceptic's two challenges. If the interpreter's projection endows

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

the agent's preferences with a dimension of strength that possesses the same general characteristics as the interpreter's preferences, then, according to Davidson, preference strengths are formed and operate in interpersonally similar ways. If this is the case, it follows that there do exist positive comparative relations holding between different individuals' preference strengths. The sceptic's second challenge is met: preferences are interpersonally comparable in terms of strength. In turn, the interpersonal comparability of preferences provides the ground for making explicit ICs at a later stage. The idea is that, if preference strengths are formed and operate in interpersonally similar ways, then, if two individuals are in causally identical circumstances and show identical behaviour, their preference strengths are also the same.[5] This means that it is possible (at least in principle) to have knowledge of how different people's preferences compare in terms of strength. The sceptic's first challenge is met too.

Before proceeding, let me point out, for the sake of clarity, that, while the interpreter projects similar preference strengths onto the interpreted agent, it does not follow, as an effect of such a projection, that the interpreter and the interpreted agent's preference strengths can be represented on an *absolute* utility scale, i.e. by a utility function that is absolutely unique. Indeed, preference strengths are still measured on an interval scale.[6] As such, utility functions that are affine transformations of one another will do equally well, i.e. they will preserve the same information about preference strengths. The upshot of Davidson's argument is rather that, as an effect of the interpreter's projection, the interpreter's and the interpreted agent's preference strengths can be represented on a *common* interval (or ratio) utility scale. This is enough for having interpersonal comparability. One related consequence is that Davidson's argument establishes, at most, the possibility of ICs of *relative* utilities, i.e. comparisons concerning people's utility for one prospect relative to another, not the possibility of ICs of *absolute* utilities.[7] With these clarifications in mind, we are now ready for examining Davidson's argument in more detail.[8]

[5] This raises the questions of what the relevant causes and behavioural manifestations of preferences are. Neither question has received the attention that it deserves. Typically, economists take choice behaviour to be the sole relevant output of preferences. However, other kinds of behavioural manifestations might be considered relevant, such as the individuals' latency of choice, i.e. the time delay between the presentation of the option and the actual choice-making (Waldner 1972), their verbal (Harsanyi 1955) and facial (List 2003) expressions, their expressive reactions (Weintraub 1998), etc.

[6] Or a ratio scale, if, as some authors maintain (e.g. Bradley 2007), preference strengths admit of a natural zero.

[7] I would like to thank an anonymous referee for inviting me to draw attention to this point.

[8] It is also worth pointing out that Davidson's argument does not entail any specific conclusion about the third step of the original framework, i.e. the step concerned with

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

## 4. A 'STRONG' TRANSCENDENTAL ARGUMENT

In this section, I will try to reconstruct the logical structure of Davidson's argument and to discuss some objections against it. Davidson's argument seems to be based on two main premises. The first is that *correct* interpretation of other people's behaviour is possible. The second is that the PoC is 'a *necessary* condition of correct interpretation' (Davidson 2004: 72). The PoC is indeed the *deus ex machina* in Davidson's argument. It is the PoC that recommends the interpreter's projection that allegedly renders preferences interpersonally comparable in terms of strength from the start. According to Davidson, the second premise is not 'an empirical claim or an assumption for the sake of science' (Davidson 2004: 73). The PoC is neither discovered, nor pragmatically chosen, but it is an a priori principle. Thus, Davidson offers an a priori argument for thinking that preferences are interpersonally comparable in terms of strength. More specifically, given the necessary character of the PoC for correct interpretation, Davidson's a priori argument takes the form of a 'strong' transcendental argument for the interpersonal comparability of preference strengths.

The general idea behind a 'strong' transcendental argument is that we can reason from the fact that we possess a cognitive ability, first, to the individuation of the conditions that necessarily support our ability and, second, to the truth of those conditions. The standard argument goes as follows:

(I) We possess the ability to Φ;
(II) We could not Φ without *X*;
(III) Hence, *X* is true.

In the case under consideration, Φ is 'the ability to correctly interpret each other', while *X* is 'the PoC'. The argument becomes:

(1) We possess the ability to correctly interpret each other;
(2) We could not interpret each other correctly without the PoC;
(3) Hence, the PoC is true.

This argument can be used to defend the thesis that preference strengths are interpersonally comparable in virtue of the following additional claims:

> the formulation of a normative judgement. Its aim is simply to show that the basis for determining the strength of people's preferences and for making ICs of preference strength is the same. The (important) question of how we should use such comparisons for normative purposes is left entirely open.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

(4) The PoC implies that the interpreter's and the interpreted agent's preferences are comparable in terms of strength;
(5) Therefore, preferences are interpersonally comparable in terms of strength.

Our next step consists in examining whether or not the premises of Davidson's transcendental argument are sound. Given the purpose of this paper, it is worth beginning by considering whether Davidson's conclusion about preference strengths really follows from his argument in favour of the PoC. That is, it is worth beginning our assessment from premise (4). Suppose, for the moment, that the PoC is true. The question is the following: Is it really the case that the PoC implies the comparability of different people's preferences? At first sight, the answer is trivially affirmative. The PoC requires the interpreter to project certain aspects of her own mental states onto the interpreted agent. The projection makes the interpreter's and the agent's mental states similar with respect to these aspects. In turn, interpersonal similarity assures that the interpreter's and the agent's mental states are interpersonally comparable. Applied to the present case, this reasoning seems to support the conclusion that the PoC implies the interpersonal comparability of preferences in terms of strength.

There is, however, room for disagreement. For instance, Ruth Weintraub (1998: 308–310) claims that the soundness of premise (4) depends on whether the property of being interpersonally comparable in terms of strength is really one of the aspects that the interpreter is required to project onto the other agent. Weintraub does not think so. If she is right, the implications for the problem of ICs of preference strength are serious. Davidson's argument would support the conclusion that the interpreter's and the agent's mental states are interpersonally comparable with respect to *some* dimensions, but not with respect to *strength*. Weintraub argues as follows. The PoC requires the interpreter to project only those mental states and properties whose existence she has reason to believe in. Indeed, according to Weintraub, the PoC does not require the interpreter to maximize similarities between individuals, e.g. by multiplying the mental states and properties to be projected onto the interpreted agents, but rather to minimize differences.[9] On the other hand, the interpreter has reason to believe in the existence of only those mental states and properties that play a role in the explanation and for the guidance of her own behaviour. The problem is that the property of being interpersonally comparable in

[9] While, strictly speaking, Weintraub's remark about the PoC is misleading, in the sense that the PoC enjoins us neither to maximize similarities, nor to minimize differences, but rather, in Davidson's own words, to 'optimize agreement' (Davidson 1984: 137), this point does not affect Weintraub's argument.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

terms of strength, which preferences might have, does not contribute in any way to directing or explaining the interpreter's behaviour. Therefore, it can be dispensed with as gratuitous and unnecessary. It follows that the interpersonal comparability of preference strengths is not one of the aspects that the interpreter is required to project onto the other agent. As such, premise (4) appears to be false: interpersonal comparability is not implied by the principle that governs correct interpretation.

At this stage, there seem to be only two possibilities left. The first is to argue that the interpersonal comparability of preference strengths does contribute to the explanation or the guidance of (at least some aspects of) the interpreter's own behaviour. If this is true, then the PoC will indeed recommend the interpreter to project this property onto the interpreted agent, as stated by premise (4) in Davidson's argument. While promising in principle, however, this line of thought is highly controversial – as, incidentally, Davidson himself recognizes (Davidson 2004: 64–67). So, I shall not explore it any further here. The second possibility consists in granting the fact that the interpersonal comparability of preference strengths does not add anything to the explanation and guidance of the interpreter's behaviour, while arguing that there are nonetheless other reasons for projecting this property onto the interpreted agent. For instance, one may try to pursue a popular strategy in the literature and appeal to pragmatic considerations (Harsanyi 1955, 1977; Waldner 1972; List 2003). The thought is that, in the present case, the interpreter has the choice of projecting either interpersonal comparability or interpersonal incomparability of preference strengths onto the interpreted agent. Yet, the former seems to be the simplest, the most parsimonious and the least arbitrary of the two options. Hence, it should be favoured on such pragmatic grounds. Unfortunately, this line of thought will not work either as a response to Weintraub's objection. The main problem is that, from the fact that the interpreter is not required to project interpersonal comparability of preference strengths onto the interpreted agent, it does not follow that she is required to project interpersonal incomparability. To prove this point, it is sufficient to observe that the property of being interpersonally incomparable in terms of strength does not add anything to the explanation and guidance of the interpreter's own behaviour. As such, the interpreter has no more reason to project this property onto the interpreted agent than she has to project interpersonal comparability. The interpreter is not required to project either property. Therefore, the appeal to pragmatic considerations is simply beside the point.

I have claimed that Weintraub's objection seems to leave us with only two possibilities. However, both of them fail. Shall we conclude that Davidson's argument is unsuccessful? I think that this would be premature. There is in fact another, largely neglected possibility to counter Weintraub's objection that needs to be taken into account. The underlying

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

idea is that it is possible *both* to accept Weintraub's main point *and* to defend the soundness of premise (4). To see why, notice, first, that, while it might well be true that the ascription of *interpersonally comparable* preference strengths does not add anything to the explanation of the interpreter's own behaviour, it does not follow that the ascription of preference *strengths* is likewise unnecessary. Quite the contrary, reference to preference strength is relevant to explain at least some aspects of the interpreter's behaviour (cf. Harsanyi 1955; Davidson 2004). Here is a quick example. Suppose that an individual is willing to exchange a lottery $A$ in which the prizes $y$ and $w$ occur with probability 0.5 for a lottery $B$ in which the prizes $x$ and $z$ occur with probability 0.5. In this case, we can clearly explain the individual's behaviour by saying that she prefers $x$ to $y$ *more strongly* than she prefers $w$ to $z$.[10] That is, we can explain her behaviour in terms of preference strengths. Consider now the implications that this point has for our discussion. If strength is a dimension of preferences that contribute to the explanation of the interpreter's behaviour, then the PoC will require the interpreter to project this aspect onto the other agent. But if the interpreter projects a dimension of strength onto the agent's preferences, then it follows that the interpreter's and the agent's preferences are similar with respect to strength. And, as we have seen in section 2, this guarantees that they are also interpersonally comparable in terms of strength. The crucial thing to notice is that the interpersonal comparability of preference strengths is not *projected* onto the other agent, but *follows* from the interpreter's projection. This point is implicit in Davidson's discussion, but has not been adequately recognized in the literature. Yet, it needs emphasis, not the least because it permits us to accept the central point of Weintraub's objection without also having to accept her conclusion. Indeed, we can accept, as Weintraub maintains, that the property of being interpersonally comparable does not add anything to the explanation of the interpreter's own behaviour and that the interpreter is thereby not required to project it onto the interpreted agent. This notwithstanding, we can reject the conclusion of Weintraub's argument by recognizing that the property of being interpersonally comparable follows from the projection of another, explanatory relevant property, namely, the property of the interpreter's preferences of having a dimension of strength. Therefore, Weintraub's objection fails: the interpersonal comparability of preference strengths is indeed implied by the principle that makes correct interpretation possible, as premise (4) states. In other words, *if* the first part of Davidson's argument is correct and the PoC is true, *then* premise (4) is sound. The

[10] Indeed, this is von Neumann and Morgenstern's proposed definition of preference intensity (von Neumann and Morgenstern 1944: 18; cited in Binmore 2009: 546).

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

question is whether the first part of Davidson's argument is really correct. To this, I shall now turn.

Let us consider premise (1). One of the salient features of transcendental arguments is that the first premise states certain facts that even the sceptic can grant. However, the soundness of (1) is not immediately obvious. There is no doubt that we can interpret each other's behaviour. However, (1) holds more stringently that we can interpret each other's behaviour *correctly*. This is less obvious (to say the least). It is also unclear whether, and how, we can assess the soundness of (1) a priori, independently from (2), that is, without knowing the conditions that are necessary for correct interpretation (cf. Lepore and Ludwig 2005: 204). The sceptic might perhaps concede that (1) is sound if we conceive correctness in the weaker sense of rational intersubjective agreement. The problem is that, while rational intersubjective agreement offers perhaps a good reason for believing that an interpretation is correct, it is no proof that such an interpretation is really correct in the required sense. In a similar vein, the sceptic may grant that we possess the ability to *rationally* interpret each other. Once again, however, rational interpretation is not synonymous with correct interpretation. Therefore, premise (1) remains dubious.

Let us consider (2). The first question that we need to address is how to precisely define the PoC. This task is surprisingly difficult. Davidson himself has given conflicting accounts of the PoC (Davidson 1984, 2003, 2004) and finding a consistent definition has proven hard for all the authors that have confronted themselves with this task (e.g. Lepore and Ludwig 2005: chapter 12). Roughly speaking, as we have seen, the PoC requires the interpreter to favour the interpretation that maximizes the number of truths that the interpreted agent believes (Davidson 1984). If this characterization is correct, however, the PoC seems to set too strong a requirement on interpretation (see, for instance, the objections raised by Grandy 1973; McGinn 1977; Dennett 1987; Lepore and Ludwig 2005). For illustrative purposes, consider the following example from the literature:

Suppose Paul has just arrived at a party and asserts: 'The man with a martini is a philosopher.' And suppose that the facts are that there is a man in plain view who is drinking water from a martini glass and that he is not a philosopher. Suppose also that in fact there is only one man at the party drinking a martini, that he is a philosopher, and that he is out of sight (Grandy 1973: 445).

If the interpreter follows the PoC's recommendation to make the agent's beliefs as true as possible, then she will ascribe to Paul the belief that the man out of sight is a martini-drinking philosopher. By so doing, however, the interpreter gets things wrong. Intuitively, the correct interpretation is the one that ascribes to Paul the false belief that the man in plain view is a martini-drinking philosopher. This shows that the PoC

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

is not an adequate basis for correct interpretation. The soundness of (2) is therefore questionable.[11]

Quite independently from how the PoC is characterized, premise (2) is also vulnerable to Barry Stroud's objection against *all* transcendental arguments of a 'strong' form (Stroud 1968). In a nutshell, Stroud's idea is that 'strong' transcendental arguments do not succeed either because they require the sceptic to make commitments that she may not be willing to make or because they are unable to obtain a conclusion about a mind-independent world from premises about our psychological reality or capacities. Recall the general form of a 'strong' transcendental argument. (II) claims that we could not $\Phi$ without $X$, where $\Phi$ is typically a fact about thought or experience and $X$ is a mind-independent fact about the world. The sceptic may object that the possibility of $\Phi$ing merely requires that we *believe* $X$, but that $X$ need not be actually true. The conclusion of a transcendental argument is then a conclusion about our beliefs, rather than about the world. The proponent of the transcendental argument has two possibilities. Either she attempts to bridge the gap between beliefs and reality by claiming that, in order to meaningfully believe $X$, we must effectively know that $X$ is true; or she collapses the belief/reality divide by arguing that reality is, at least in part, constituted by how we take things to be. However, neither possibility is very appealing. The first invokes a verificationist theory of meaning that has been discredited a long time ago and that would make the transcendental argument entirely redundant.

[11] As one referee has pointed out to me, Grandy's attack against the PoC rests perhaps on too narrow a view of the conditions for belief ascription that the PoC imposes. A broader view would maintain that the PoC requires the interpreter to maximize the coherence between actions and beliefs, as well as the truth of beliefs. If we adopt this view, then Paul's behaviour is best explained by attributing to him the false belief that the man in plain view is a martini-drinking philosopher, since his further behaviour is likely to prove incoherent with an alternative ascription. This seems to me correct. However, this strategy opens the door to another objection against the PoC. Truth-maximization is now only one of the criteria for mental ascription, the other being coherence-maximization. If this is the case, the formulation of the PoC should be refined by specifying the relative importance of the different criteria of mental ascription. Unfortunately, Davidson's remarks on the subject are quite elusive. Even so, there are reasons to think that this strategy would not work. The thing is that, if truth-maximization turns out to be the dominant principle, then it is possible to argue against the PoC by means of examples analogous to the one suggested by Grandy. On the other hand, if truth-maximization is assigned a lower importance than coherence-maximization, or if the interpreter's goal is simply to 'optimize agreement' between herself and the interpreted agent, then there is no guarantee that the PoC will favour the interpretation that makes *most* of the agent's beliefs true, as Davidson suggests. Of course, Davidson might downplay the role of truth for belief ascription and emphasize the role of interpersonal similarity. But this would make the characterization of the PoC closer and closer to the Principle of Similarity, which I shall present below, and with that characterization I have no qualms. For similar criticisms of the PoC, see, amongst others, McGinn (1977), Lepore and Ludwig (2005) and Goldman (2006).

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

The second leads to embracing an idealist position, which looks more like a form of scepticism than a refutation of it (cf. Stern 2000: chapter 2).

Stroud's objection applies to the case under consideration as well. Suppose, for the sake of the argument, that we can correctly interpret each other and that there exists a version of the PoC that is immune to counter-examples. The sceptic may still object that the possibility of correct interpretation requires only that we believe the PoC, not that the PoC be really true. But then Davidson's conclusion that preferences are interpersonally comparable in terms of strength does not follow. We cannot infer the conclusion that preferences are *really* interpersonally comparable simply from a belief about the PoC. At best, the belief that the PoC is true implies the *belief* that preferences are interpersonally comparable. However, this belief might be false. Once again, the proponent of the transcendental argument has the option of adding either a verificationist or an idealist premise in order to meet the sceptic's challenge. As we have seen, neither option is particularly attractive. If Stroud's general objection is correct, then Davidson's proposal does not succeed.

## 5. A 'MODEST' TRANSCENDENTAL ARGUMENT

Despite the fact that 'strong' transcendental arguments face serious problems, Stroud himself and many others have argued that more 'modest' transcendental arguments may still achieve results of anti-sceptical significance (see Stern 1999; Stroud 2002). For the present purpose, we shall consider three questions. First, how can Davidson's argument be transformed into a 'modest' transcendental argument? Second, can the resulting argument be defended? Third, what are the implications for the problem of ICs of preference strength? I shall attempt to answer the first two questions in this section, while I shall reserve a discussion of the third one for the next section.

The objections presented in the previous section challenge both premise (1) and premise (2) of Davidson's argument. The objection against (1) is that we can only believe that correct interpretation is possible, without being able to know that it really is. The first objection against (2) is that the PoC does not provide an adequate basis for interpretation. The second objection against (2) is that, whatever the right principle governing interpretation, it is not possible to know that such a principle is necessary for the correct interpretation of other people's behaviour, but only to believe that it is.

The first step to transform Davidson's 'strong' transcendental argument into a convincing 'modest' transcendental argument consists in finding a more appropriate basis for interpretation than the PoC. What we need is not a principle that makes the interpreted agent's beliefs as true

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

as possible, but a principle that makes the most *sense* of the interpreted agent's mental states and behaviour, by the interpreter's own light. Many authors have argued that the required principle is the Principle of Similarity (PoS henceforth) (Grandy 1973; Dennet 1987; Goldman 1989).[12] Indeed, the PoS is sufficient to capture Davidson's idea that 'similar causes beget similar evaluations in interpreter and interpreted' (Davidson 2004: 72), which supposedly is at the heart of the interpretative procedure. We can characterize the PoS as follows:

*The Principle of Similarity*: for any two individuals, *ceteris paribus*, if they were subject to the same circumstances, then they would form the same beliefs, desires, preferences, intentions, etc.

My proposal is that, if we replace the PoC with the PoS and if we modify premises (1) and (2) in such a way to take the previous objections into account, then we can reformulate Davidson's argument in the following, more 'modest' way:

(1′) We can rationally interpret other people's behaviour;
(2′) We could not rationally interpret other people's behaviour without believing the PoS;
(3′) Hence, we believe the PoS.
(4′) Believing the PoS implies believing that preferences are interpersonally comparable in terms of strength;
(5′) Therefore, we believe that preferences are interpersonally comparable in terms of strength.[13,14]

Let us now proceed to the assessment of premises (1′)–(4′) of this 'modest' transcendental argument, keeping in mind that the inference from (4′) to (5′) will be discussed in detail in the next section. Consider (1′). As we have seen above, it seems plausible to hold that we can rationally interpret other people's behaviour. This is a claim that even the sceptic may concede. Consider (4′) now. Even though Davidson's 'strong' transcendental argument fails, the points made while discussing Weintraub's objection apply here, *mutatis mutandis*. The belief that people

[12] Grandy refers to this principle as the 'Principle of Humanity'.

[13] While I am formulating the 'modest' transcendental argument in terms of the attitude of 'believing', it should be clear that I am using this notion broadly, to refer to any of the doxastic attitudes (e.g. 'taking', 'seeming', 'appearing', etc.) that can most plausibly make the argument successful.

[14] One may object that this argument is too different from Davidson's original argument to count as properly 'Davidsonian'. The point can be granted. In this paper, however, I am less concerned with a careful exegesis and defence of Davidson's own position than with the exploration of the implications that his position has for the problem of ICs, once it is modified in a plausible way. My goal is to give Davidson's *strategy*, rather than Davidson's particular *view*, its best shot.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

respond to the same overall circumstances and form preferences in the same way implies the belief that their preferences possess the same, or similar, general characteristics. In turn, this belief implies the further belief that it is possible to establish positive comparative relations between preference strengths and, thereby, that preferences are interpersonally comparable in terms of strength. So, the soundness of premise (4′) can also be granted relatively easily.

The crucial premise is (2′). According to (2′), believing the PoS is indispensible for being able to rationally interpret other people's behaviour. Given the previous definition of the PoS, (2′) is false if we can rationally interpret an agent's behaviour, but maintain that, *ceteris paribus*, we would not form our mental states as the agent did, were we subject to identical circumstances. Clearly, the latter belief must still hold once *all* the circumstances that are deemed relevant for the agent's behaviour have been taken into account. The sceptic may suggest that this possibility is perfectly real. One way to make this possibility vivid is by considering the following thought experiment.

Suppose that an interpreter *a* believes herself to be radically different from another individual *i*. Differences between interpreter and interpreted agent may concern all the dimensions required for interpretation, namely, the function, content and strength of each mental state and the relation between different mental states. Suppose, however, that the interpreter manages to interpret *i*'s behaviour since she possesses a book about *i*'s life, which reliably predicts *i*'s behaviour and indicates the mental states that *i* will form in each specific circumstance of her life. In order to defend (2′), one needs to deny that this scenario constitutes a genuine possibility. Such an interpretation seems to be rational, since it is based on reliable evidence about the individual *i*'s behaviour, i.e. the book about *i*'s life. Thus, the reason for dismissing the sceptic's challenge must be another. One possibility consists in denying that the interpreter can provide an *intelligible* explanation of *i*'s behaviour. Recall that, according to Davidson, differences between interpreter and interpreted agent can be tolerated up to the point where the agent's behaviour remains intelligible. That is, the explanation of the agent's behaviour must be intelligible in order to preserve understanding. If intelligibility is a necessary condition for rational interpretation, then the sceptic's case is unsuccessful.

Let me spell out the argument in favour of this conclusion in more detail. Intelligibility requires that the concepts on which that interpretation is based be appropriately applied, *according to the interpreter's own standards*. The question concerns what these standards are and how they are acquired. Here, we can make appeal to a slightly modified version of Davidson's 'triangulation' strategy (Davidson 2003). According to Davidson, an interpreter learns how to apply mental concepts by correlating mental events with external circumstances and behaviour. The learning process typically involves a triangulation

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

between two (or more) persons and the world, to the effect that each person learns how to apply mental concepts by observing both the other individual's and her own reactions to identical, or similar, environmental stimuli. Through triangulation, then, different people come to associate the same mental concepts to the same, or relevantly similar, circumstances and to believe that these concepts are correctly applied in those circumstances. But this means that each person comes to embrace the PoS and to think that, *ceteris paribus*, if two individuals are subject to the same circumstances, they form the same mental states. This has some important implications for the issue under consideration. If mental concepts are used in too different a way, the interpreter will not take them to be appropriately applied and, thereby, intelligibility will not be possible. Instead, the interpreter will find the explanation of the agent's behaviour intelligible only provided that the mental concepts on which it is based are applied according to standards that the interpreter recognizes as holding in her own case as well as in other individuals' case. As we have seen, this condition stems from the interpreter's commitment to the PoS, as developed through triangulation. If this is the case, then it is impossible for the interpreter to find the explanation of another agent's behaviour intelligible (by her own light) and yet not to believe the PoS. This shows that the sceptic's thought experiment is not a genuine counter-example to (2′).

## 6. THE PROBLEM OF INTERPERSONAL COMPARISONS RECONSIDERED

Let us take stock and summarize the results reached so far. We have seen that Davidson's original argument does not succeed since it falls prey to several objections, including Stroud's powerful objection against all transcendental arguments of a 'strong' kind. However, Davidson's argument can be turned into a successful transcendental argument of a 'modest' kind. The next question on our agenda concerns the implications that these results have for the epistemological issues concerning ICs of preference strength.

In order to address this question, I shall introduce a distinction between two forms of scepticism, epitomized by the figures of the 'epistemic sceptic' and the 'justificatory sceptic' (see Stern 2000: chapter 1, for this distinction). The 'epistemic sceptic' is the (possibly fictitious) individual who doubts whether we can have *knowledge* of specific subject matters.[15] Instead, the 'justificatory sceptic' is the (possibly

[15] In this sense, the label of 'knowledge sceptic' would perhaps be preferable. For simplicity, however, I will stick to Stern's original label.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

fictitious) individual who doubts whether we can have *justified beliefs* about these subject matters. According to Stern, 'strong' transcendental arguments are directed towards the 'epistemic sceptic', whereas 'modest' transcendental arguments are directed towards the 'justificatory sceptic'. As a consequence, the failure of 'strong' transcendental arguments compromises only the possibility of having genuine knowledge of specific subject matters, but not the possibility of having justified beliefs about them. In the case under consideration, Stern's observations suggest that, although we cannot be certain of how different people's preferences compare in terms of strength and, ultimately, of whether preferences are really interpersonally comparable in terms of strength, nonetheless, given the success of the 'modest' transcendental argument, we can have justified beliefs about these matters.

At first sight, this claim may seem puzzling. Premise (4') in the 'modest' transcendental argument assures a transcendental invulnerability to the belief that preference strengths are interpersonally comparable. However, the fact that we cannot but assume interpersonal comparability in order to rationally interpret each other seems to fall short of what we need to address the concerns of *all* sorts of sceptic. What we are bound to do is in fact a descriptive matter, while knowledge and justification are evaluative matters. Clearly, the fact that we cannot but take preferences to be interpersonally comparable does not entail that preferences are really interpersonally comparable in terms of strength. That is, as expected, the 'modest' transcendental argument is unable to satisfy the demands of the 'epistemic sceptic'. But the fact that the belief about the interpersonal comparability of preference strengths is transcendentally invulnerable does not entail that such a belief is justified either. That is, and more surprisingly, the 'modest' transcendental argument seems to be unable to address 'justificatory scepticism' about ICs of preference strength as well!

Let us consider an example to make this point vivid. Suppose that an interpreter *a* believes, on the basis of the empirical evidence, that (i) *i*'s and *j*'s relevant circumstances are the same. Suppose that she forms an interpretation of *i*'s and *j*'s behaviour on the basis of such a belief. The 'modest' transcendental argument shows that, in order to rationally interpret *i*'s and *j*'s behaviour, *a* must believe that (ii) for any two individuals *i* and *j*, *ceteris paribus*, if their circumstances were the same, they would form the same mental states. In the present case, if *a* does interpret *i*'s and *j*'s behaviour rationally, then she will believe that (iii) *i*'s and *j*'s mental states are the same. This follows in fact from the combination of (i) and (ii). So much can be granted. Nevertheless, the 'justificatory sceptic' may remain unimpressed. She can concede that it is possible for *a* to believe (i)–(iii), while denying that it is possible for her to make any *justified* ICs of *i*'s and *j*'s preference strengths and to have any *justified* beliefs about their interpersonal comparability.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

Against the 'justificatory sceptic', I want to suggest that, if we combine the 'modest' transcendental argument with an 'internal' account of epistemic justification, according to which epistemic justification derives from the satisfaction of 'our deepest logical and intellectual standards and procedures' (Sosa 1988: 183), *then* it is indeed possible to successfully address both of the justificatory sceptic's challenges. Consider the first challenge. If, more specifically, the 'internal' standards of epistemic justification require the interpreter to form her beliefs on the basis of the empirical evidence in a consistent and coherent way,[16] then the 'modest' transcendental argument shows that the interpreter cannot do any better, epistemically, than forming some specific beliefs about how different people's preferences compare in terms of strength. Such beliefs uniquely satisfy the 'internal' requirements for epistemic justification. Therefore, they are epistemically justified. This means that, *pace* the 'justificatory sceptic', it is possible for an interpreter to make justified ICs of preference strength. Indeed, the conclusion is even stronger. Not only is it possible for an interpreter to make justified ICs of preference strength; but, given the transcendental indispensability of the interpreter's belief about the PoS, it is necessarily true that if an interpreter can rationally interpret other individuals' behaviour, then it is possible for her to make epistemically justified ICs of their preference strengths.

For illustrative purposes, let us reconsider the example above. The previous discussion shows that believing that (ii) for any two individuals *i* and *j*, *ceteris paribus*, if their circumstances were the same, they would form the same mental states implies believing that (iv) for any two individuals *i* and *j*, *ceteris paribus*, if their circumstances were the same, they would form the same preference strengths. This has the following implications. First, given (i) and (iv), *a*'s belief-set is more coherent if she possesses the belief that (v) *i* and *j* have the same preference strengths. Thus, if epistemic justification requires the interpreter to form coherent beliefs on the basis of the relevant evidence, then (v) satisfies at least one of the conditions for counting as an epistemically justified belief. The sceptic may insist that even if *a* coherently believes (v), it may still be the case that (v') *i* and *j do not* have the same preference strengths. This is indeed a logical possibility. Yet, there is a sense in which 'justificatory scepticism' does not get off the ground. Suppose in fact that the sceptic persuades the interpreter to believe (v'). Her belief-set looks now inconsistent. If epistemic justification requires the interpreter to revise her beliefs in the face of recognized inconsistencies, then we obtain an interesting result. While, in principle, the interpreter can achieve consistency by revising any of the beliefs in her belief-set, the situation looks very different if the

[16] This 'internal' account of epistemic justification would thus combine both evidentialist (see Feldman and Conee 1985) and coherentist (see Dancy 1985) elements.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

'modest' transcendental argument seen above is successful. Indeed, given that (iv) inherits the transcendental invulnerability of (ii), and assuming, as it seems plausible, that (i) can be justified on independent grounds, the interpreter has only one choice available to make her belief-set consistent: she must abandon (v′) and adopt (v) instead. In other words, (v) is the *only* belief about *i*'s and *j*'s preference strengths that can preserve consistency, given the available empirical evidence. The interpreter cannot believe (i)–(iii) and consistently form a belief about *i*'s and *j*'s preference strengths other than (v).[17] In accordance with the 'internal' account of epistemic justification, (v) is thus epistemically justified.

The conclusion is the following. Although it is not possible to show that we can have knowledge of ICs, it is nonetheless possible to show that we can have epistemically justified comparisons of different people's preference strengths, provided that we adopt an 'internal' account of epistemic justification. Not only; but the 'modest' transcendental argument shows also that, necessarily, if we can rationally interpret other people's behaviour, then we can also make justified ICs of their preference strengths. Finally, if the previous analysis is correct, it is possible to say something against the 'justificatory sceptic' about the interpersonal comparability of preference strengths as well. Here is a very simple argument, which draws on the previous considerations. If specific ICs of preference strength are justified, then we are justified in believing that it

[17] Here is a possible objection, presented to me by Richard Bradley (personal communication). The sceptic may reject this argument by attacking the claim that (i) can be justified on independent grounds. For instance, the sceptic may argue as follows. The PoS is plausible only if we interpret the notion of 'relevant circumstances' broadly enough to include the individuals' constitutions, in addition to the individuals' external circumstances. However, it is impossible to justifiably believe that two individuals' constitutions are *really* the same, without having justified beliefs about how their preferences compare in terms of strength. But this implies that, in order to justifiably believe that two individuals' overall circumstances are really the same, one should already know how to solve the problem of ICs. If this is true, the argument offered above is question-begging, because it simply presupposes a solution to the problem of ICs. The objection is interesting and deserves further scrutiny. For reasons of space, my response here will be brief. One possibility is clearly to give a non-psychological (e.g. biological, neurophysiological, etc.) characterization of the notion of 'constitution'. In this case the problem is entirely avoided. But even if the characterization of 'constitution' is, in some respect, psychological, there is some room for a reply. The starting point is the observation that the problem of ICs is not a form of global scepticism about the mind, according to which it is completely impossible to have justified beliefs about any aspect whatsoever of the individual's mind. So, *unless* we define an individual's constitution as the set of dispositions to form mental states with a specific strength – in which case the PoS would essentially amount to a tautology – then it may be possible to have justified beliefs about different individuals' constitution without having prior justified beliefs about their preference strengths. If this is the case, it may be possible to justify (i) independently, without presupposing a solution to the problem of ICs.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

is possible to establish positive comparative relations between different people's preference strengths. That is, we are justified in believing that preferences are interpersonally comparable in terms of strength. More strongly, the 'modest' transcendental argument shows that, necessarily, if we can rationally interpret other people's behaviour, then we can also have justified beliefs about the comparability of their preference strengths. By addressing the 'justificatory sceptic' about specific ICs of preference strength, we can dismiss the 'justificatory sceptic' about the interpersonal comparability of preference strengths. Both of the challenges raised by the 'justificatory sceptic' are met.

## REFERENCES

Arrow, K. 1963. *Social Choice and Individual Values*. New York: Wiley.
Binmore, K. 2009. Interpersonal comparison of utility. In *The Oxford Handbook of Philosophy of Economics*, eds. H. Kincaid and D. Ross, 540–559. Oxford: Oxford University Press.
Bradley, R. 2007. Comparing evaluations. *Proceedings of the Aristotelian Society* 108: 85–100.
Dancy, J. 1985. *An Introduction to Contemporary Epistemology*. Oxford: Basil Blackwell.
Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.
Davidson, D. 1986. Judging interpersonal interests. In *Foundations of Social Choice Theory*, eds. J. Elster and A. Hylland, 195–211. Cambridge: Cambridge University Press.
Davidson, D. 2003. *Subjective, Intersubjective, Objective*. Oxford: Oxford University Press.
Davidson, D. 2004. *Problems of Rationality*. Oxford: Oxford University Press.
Dennett, D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press.
Feldman, R. and E. Conee 1985. Evidentialism. *Philosophical Studies* 48: 15–34.
Goldman, A. 1989. Interpretation psychologised. *Mind and Language* 4: 161–185.
Goldman, A. 2006. *Simulating Minds*. Oxford: Oxford University Press.
Grandy, R. 1973. Reference, meaning, and belief. *Journal of Philosophy* 70: 439–452.
Griffin, J. 1986. *Well-Being: Its Meaning, Measurement, and Moral Importance*. Oxford: Oxford University Press.
Hammond, P. 1991. Interpersonal comparisons of utility: why and how they are and should be made. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. Roemer, 200–254. Cambridge: Cambridge University Press.
Harsanyi, J. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy* 63: 309–321.
Harsanyi, J. 1977. *Rational Behaviour and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.
Hausman, D. 1995. The impossibility of interpersonal utility comparisons. *Mind* 104: 473–490.
Isbell, J. 1959. Absolute games. In *Contributions to the Theory of Games*, ed. A.W. Tucker and R.D. Luce, 357–396. Princeton: Princeton University Press.
Jeffrey, R. 1971. On interpersonal utility theory. *Journal of Philosophy* 68: 647–656.
List, L. 2003. Are interpersonal comparisons of utility indeterminate? *Erkenntnis* 58: 229–260.
Lepore, E. and K. Ludwig 2005. *Donald Davidson. Meaning, Truth, Language, and Reality*. Oxford: Clarendon Press.
McGinn, C. 1977. Charity, interpretation, and belief. *Journal of Philosophy* 74: 521–535.
Ramsey, F. 1926. Truth and probability. Reprinted in *Philosophical Papers*, ed. D.H. Mellor, 1990. Cambridge: Cambridge University Press.
Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press

Robbins, L. 1932. *An Essay on the Nature and Significance of Economic Science*. London: Macmillan.
Schick, F. 1971. Beyond utilitarianism. *Journal of Philosophy* 68: 657–666.
Sen, A. 1970. *Collective Choice and Social Welfare*. San Francisco, CA: Holden-Day.
Sen, A. 1979. Interpersonal comparisons of welfare. In *Economics and Human Welfare: Essays in Honor of Tibor Scitovsky*, ed. M.J. Boskin, 183–201. New York: Academic Press.
Sosa, E. 1988. Beyond scepticism, to the best of our knowledge. *Mind* 97: 153–188.
Stern, R. 1999. *Transcendental Arguments: Problems and Prospects*. Oxford: Oxford University Press.
Stern, R. 2000. *Transcendental Arguments and Scepticism*. Oxford: Clarendon Press.
Stroud, B. 1968. Transcendental arguments. *Journal of Philosophy* 65: 241–256.
Stroud, B. 2002. *Understanding Human Knowledge*. Oxford: Oxford University Press.
von Neumann, J. & Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.
Waldner, I. 1972. The empirical meaningfulness of interpersonal utility comparisons. *Journal of Philosophy* 4: 87–103.
Weintraub, R. 1995. The impossibility of interpersonal utility comparisons: a critical note. *Mind* 105: 661–665.
Weintraub, R. 1998. Do utility comparisons pose a problem? *Philosophical Studies* 92: 307–319.

https://doi.org/10.1017/S0266267111000216 Published online by Cambridge University Press